Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Statistics & Data Analysis 50 (2006) 242 – 253

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Projected gradient approach to the numerical solution of the SCoTLASS

Nickolay T. Trendafilov[a,*], Ian T. Jolliffe[b]

[a] *Faculty of Computing, Engineering and Mathematical Sciences, University of the West of England, Bristol BS16 1QY, UK*

[b] *Department of Mathematical Sciences, University of Aberdeen, Aberdeen AB24 3UE, Scotland, UK*

Available online 20 August 2004

Abstract

The SCoTLASS problem—principal component analysis modified so that the components satisfy the Least Absolute Shrinkage and Selection Operator (LASSO) constraint—is reformulated as a dynamical system on the unit sphere. The LASSO inequality constraint is tackled by exterior penalty function. A globally convergent algorithm is developed based on the projected gradient approach. The algorithm is illustrated numerically and discussed on a well-known data set.

Keywords: Principal component analysis; Simple structure components; LASSO constraint; Penalty function; Constrained optimization; Gradient dynamical system on manifolds; Steepest ascent vector flows; Optimality conditions

1. Introduction

Principal component analysis (PCA) (Jolliffe, 2002) is probably the most popular descriptive multivariate technique for data analysis. Many other multivariate techniques that involve dimension reduction have links to PCA, e.g. multidimensional scaling, correspondence analysis, etc. The mathematical equivalent of PCA is the symmetric eigenvalue problem (Golub and Van Loan, 1991), for which computationally fast and stable algorithms exist (Golub and Van Loan, 1991).

[*] Corresponding author.
E-mail address: nickolay.trendafilov@uwe.ac.uk (N.T. Trendafilov).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.07.017

In standard PCA, principal component extraction is followed by some kind of transformation which aims to make the components easier for interpretation. Recently, it has been recognized that it makes more sense if the principal components extraction is performed in a way which secures their reasonable and objective interpretation (Jolliffe and Uddin, 2000). Thus the problem is to join the extraction and interpretation in a *simultaneous* procedure, e.g. to maximize the variance and the VARIMAX criterion (or some other simplicity criterion) simultaneously (Jolliffe and Uddin, 2000; Trendafilov, 2001). A new promising approach to achieve this is to impose additional constraints of Least Absolute Shrinkage and Selection Operator (LASSO) type (Osborne et al., 2000a, 2000b; Tibshirani, 1996; Turlach et al., 2001) to the principal components. Suppose we are given an $p \times p$ correlation matrix $\mathbf{R}$. The PCA constrained to fulfill the LASSO looks for $m(\leqslant p)$ vectors $\mathbf{a}_k,\ k=1,2,\ldots,m$ of size $p \times 1$ which, first, as in the standard PCA:

$$\text{Maximize} \quad \mathbf{a}_k^{\mathrm{T}}\mathbf{R}\mathbf{a}_k \tag{1}$$

$$\text{Subject to} \quad \|\mathbf{a}_k\|_2 = 1 \text{ and } \mathbf{a}_k^{\mathrm{T}}\mathbf{A}_{k-1} = (\underbrace{0,0,\ldots,0}_{k-1}) = \mathbf{0}_{k-1}^{\mathrm{T}}, \tag{2}$$

where the matrix $\mathbf{A}_{k-1}$ is composed by all preceding vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k-1}$, i.e. $\mathbf{A}_{k-1}$ is the $p \times (k-1)$ matrix defined as $\mathbf{A}_{k-1} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k-1})$.

Second, the additional LASSO condition is imposed:

$$\|\mathbf{a}_k\|_1 \leqslant t, \tag{3}$$

for some tuning parameter $t \leqslant \sqrt{p}$.

Modified PCA based on the LASSO (1)–(3) has been introduced and considered in Jolliffe et al. (2003). It is called for short SCoTLASS (Simplified Component Technique-LASSO). In order to keep the size of the paper (Jolliffe et al., 2003) reasonable, the numerical algorithm for solving the SCoTLASS problem (1)–(3) was not explained and discussed there.

This short paper is complementary to Jolliffe et al. (2003). We give detailed consideration of the numerical method for solving the SCoTLASS problem (1)–(3). The method is rather general and can be applied to a wide class of multivariate models. Although this paper is self-contained, the exposition is formal and the reader is referred to Jolliffe et al. (2003) for a more informal introduction to the SCoTLASS problem (1)–(3) and its motivation.

For every $k = 1, 2, \ldots, m$ the feasible set of the SCoTLASS problem (1)–(3) is the intersection of the following two sets:

$$\|\mathbf{a}_k\|_2 = 1 \quad \text{and} \quad \|\mathbf{a}_k\|_1 \leqslant t, \tag{4}$$

i.e. the intersection of the unit sphere with the LASSO constraint (square for $p = 2$, octahedron for $p = 3$, etc) in $\boldsymbol{R}^p$. Thus the feasible set of the SCoTLASS problem (1)–(3) is a compact not connected subset of the unit sphere in $\boldsymbol{R}^p$ with compact connected components. The PCA objective function (1) is continuous on the feasible set and, therefore, has a global maximum.

Note that the PCA objective function (1) has a global maximum on each of the connected components, with the greatest one being the global maximum for the SCoTLASS problem (1)–(3) as a whole. This fact suggests that solution of the SCoTLASS problem (1)–(3) can be found by solving the problem in each component and picking the greatest maximum.

When the tuning parameter t decreases, the surfaces of the feasible components decrease too, and can be approximated by the corresponding "hyper-chords". Thus the feasible set of the SCoTLASS problem (1)–(3) can be approximated by a compact not connected subset in $\boldsymbol{R}^p$ with compact connected *convex* components. Then the objective function (1) being convex on convex compact components attains its global maximum on the boundary of the feasible set. Thus for small values of t the SCoTLASS problem (1)–(3) can be approximately considered on the following feasible set, for every $k = 1, 2, \ldots, m$:

$$\|\mathbf{a}_k\|_2 = 1 \quad \text{and} \quad \|\mathbf{a}_k\|_1 = t. \tag{5}$$

In this paper the SCoTLASS problem (1)–(3) is reformulated as a dynamical system on the manifold defined by the constraints of the problem. For this reason it seems helpful to recall the variational definition of the PCA (Jolliffe, 2002). Additional in/equality constraints on the components can be naturally incorporated in the form of penalty functions.

2. The projected gradient approach

2.1. Basic rationale

The projected gradient approach is a specific continuous-time method based on the classical gradient approach and modified for analyzing and solving *constrained* optimization problems. It is well-known that the standard gradient approach for maximization of an objective function F is given by the following gradient dynamical system (e.g. Helmke and Moore, 1994; Hirsch and Smale, 1974):

$$\frac{\mathrm{d}\mathbf{X}(t)}{\mathrm{d}t} = \nabla F(\mathbf{X}(t)). \tag{6}$$

If $\mathbf{X}(t)$ is restricted to move on a certain feasible set the gradient $\nabla F(\mathbf{X}(t))$ in (6) may move the flow $\mathbf{X}(t)$ off the feasible set because it is determined by the function F only but not at all by the constraints imposed. The aim of the projected gradient method is to keep the flow $\mathbf{X}(t)$ "clamped" to the constraint manifold. Instead of (6) the projected gradient is concerned with the following dynamical system:

$$\frac{\mathrm{d}\mathbf{X}(t)}{\mathrm{d}t} = \pi(\nabla F(\mathbf{X}(t))), \tag{7}$$

where $\pi(\nabla F(\mathbf{X}(t)))$ is the projection of the gradient $\nabla F(\mathbf{X}(t))$ onto the tangent space of the feasible set. The flow $\mathbf{X}(t)$ defined by (7) defines a steepest ascent flow for the function F on the feasible set. See (Chu and Trendafilov, 1998) for details.

2.2. Gradient vector flows for principal component extraction based on their optimality

Suppose we are given a $p \times p$ correlation matrix $\mathbf{R}$. From the principal component optimality (Jolliffe, 2002) we know that the vector of loadings for the first principal component,

$\mathbf{a}_1$, is a vector that solves:

$$\text{Maximize} \quad \mathbf{a}^{\mathrm{T}}\mathbf{R}\mathbf{a} \tag{8}$$
$$\text{Subject to} \quad \|\mathbf{a}\|_2 = 1. \tag{9}$$

The vector of loadings for the second principal component, $\mathbf{a}_2$, is a vector that solves:

$$\text{Maximize} \quad \mathbf{a}^{\mathrm{T}}\mathbf{R}\mathbf{a} \tag{10}$$
$$\text{Subject to} \quad \|\mathbf{a}\|_2 = 1 \text{ and } \mathbf{a}^{\mathrm{T}}\mathbf{a}_1 = 0. \tag{11}$$

In general, the vector of loadings for the mth principal component, $\mathbf{a}_m$, is a vector that solves:

$$\text{Maximize} \quad \mathbf{a}^{\mathrm{T}}\mathbf{R}\mathbf{a} \tag{12}$$
$$\text{Subject to} \quad \|\mathbf{a}\|_2 = 1 \text{ and } \mathbf{a}^{\mathrm{T}}\mathbf{A}_{m-1} = \mathbf{0}_{m-1}^{\mathrm{T}}. \tag{13}$$

It can be shown that this variational definition of PCA is equivalent to m consequent ascent gradient vector flows, each of them defined on an unit sphere in $\boldsymbol{R}^p$ and orthogonal to all preceding principal components. Thus the loadings for the first $m(\leqslant p)$ principal components $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ of the $p \times p$ correlation matrix $\mathbf{R}$ can be computed as solutions of m consequent initial value problems for the following vector ordinary differential equations (ODEs):

$$\frac{\mathrm{d}\mathbf{a}_k}{\mathrm{d}t} = \pi_k \nabla_{\mathbf{a}^{\mathrm{T}}\mathbf{R}\mathbf{a}}(\mathbf{a}_k), \tag{14}$$

starting with an appropriate initial value $\mathbf{a}_{k,\mathrm{in}}$ with $\|\mathbf{a}_{k,\mathrm{in}}\|_2 = 1$ for $k = 1, 2, \ldots, m$. See Chu and Trendafilov (2001), Edelman et al. (1998) for details concerning projections on Stiefel manifold. Note that $\nabla_{\mathbf{a}^{\mathrm{T}}\mathbf{R}\mathbf{a}}(\mathbf{a}) = \mathbf{R}\mathbf{a}$ is the gradient of the function $\mathbf{a}^{\mathrm{T}}\mathbf{R}\mathbf{a}$ to be maximized with respect to the standard Frobenius (Euclidean) matrix norm (e.g. Golub and Van Loan (1991)). The projector π_k in (14) is defined as follows:

$$\pi_k = \mathbf{I}_p - \mathbf{A}_k\mathbf{A}_k^{\mathrm{T}}, \tag{15}$$

where $\mathbf{I}_p$ is an $p \times p$ identity matrix. The first principal component ($k = 1$) corresponds to the largest eigenvalue of $\mathbf{R}$, the second principal component to the next one in magnitude, and so on.

Note that by introducing a negative sign in the right-hand side of Eq. (14) we define m descent gradient vector flows for the m smallest eigenvectors of the correlation matrix $\mathbf{R}$. A matrix algorithm that performs the variational PCA has been available for many years (Brockett, 1989). It is known as the *double bracket flow* (Helmke and Moore, 1994). Since then it has been rediscovered many times in different forms and for different reasons, e.g. see Chen et al. (1998), Mahony et al. (1996) for details.

2.3. *Coping with LASSO—exterior penalty function*

Making use of the identity:

$$\|\mathbf{a}_k\|_1 = \mathbf{a}_k^{\mathrm{T}}\operatorname{sign}(\mathbf{a}_k), \tag{16}$$

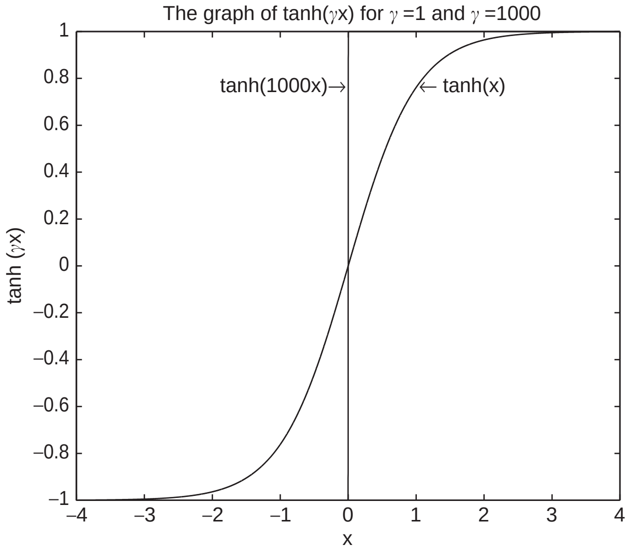


Fig. 1. Plot of $\tanh(\gamma x)$ for two γ values.

the SCoTLASS problem (1)–(3) can be rewritten in the following form (e.g. Aoki, 1971), find m vectors $\mathbf{a}_k$, $k = 1, 2, \ldots, m$ of size $p \times 1$ such that:

$$\text{Maximize} \quad F_\mu(\mathbf{a}_k) = 0.5\mathbf{a}_k^{\mathrm{T}}\mathbf{R}\mathbf{a}_k - \mu P(\mathbf{a}_k^{\mathrm{T}}\operatorname{sign}(\mathbf{a}_k) - t), \tag{17}$$

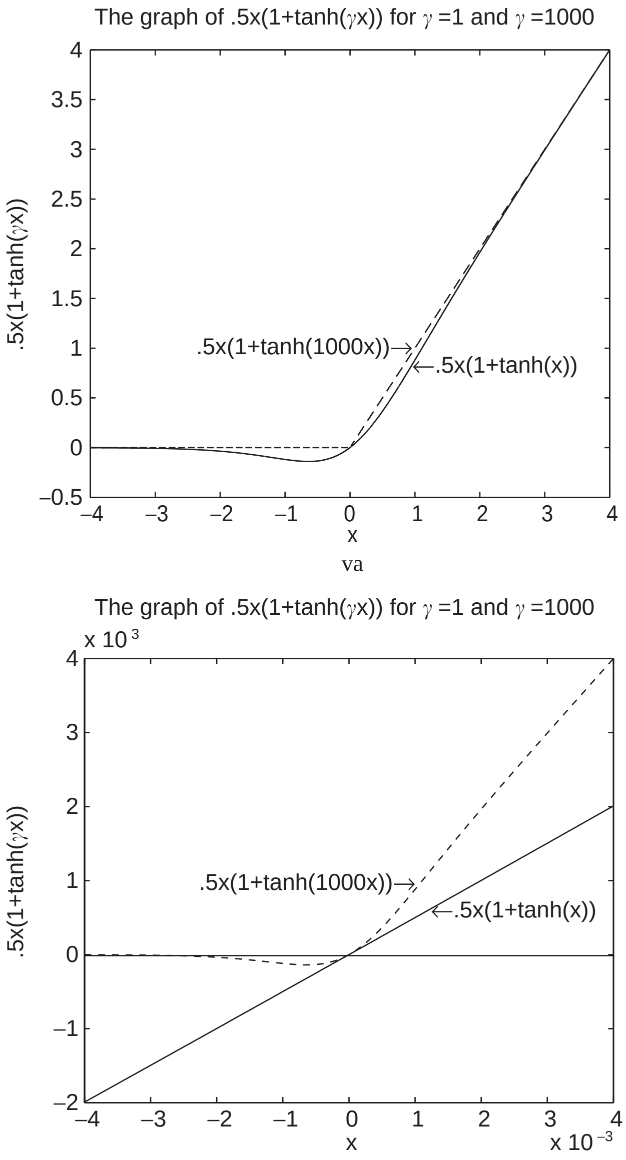
$$\text{Subject to} \quad \|\mathbf{a}_k\|_2 = 1 \text{ and } \mathbf{a}_k^{\mathrm{T}}\mathbf{A}_{k-1} = \mathbf{0}_{k-1}^{\mathrm{T}}, \tag{18}$$

where μ is some large positive number (e.g. $\mu = 1000$) and P is an exterior penalty function for inequality constraints, e.g. $P(x) = \max(0, x)$ (Zangwill penalty function) or $P(x) = \max(0, x)^2$.

The projected gradient approach requires smooth (infinitely differentiable) functions. For this reason we approximate $\operatorname{sign}(\mathbf{a}_k)$ in (17) with $\tanh(\gamma\mathbf{a}_k)$ for some sufficiently large γ, e.g. $\gamma = 1000$ in Fig. 1. See Osborne et al. (2000b) for other smooth approximations.

The standard exterior penalty functions mentioned above can be replaced by the following smooth functions, respectively: $P(x)=0.5x(1+\tanh(\gamma x))$ and $P(x)=[0.5x(1+\tanh(\gamma x))]^2$. In this paper we employ $P(x)=0.5x(1+\tanh(\gamma x))$ depicted in (Fig. 2) for two reasons. The numerical algorithm employing the quadratic penalty function $P(x)=[0.5x(1+\tanh(\gamma x))]^2$ requires more CPU time and produces loadings for which the LASSO operator takes values slightly above the required tuning parameter t which in turn leads to overestimated maximum of the objective function (1).

The left-hand side plot in Fig. 2 shows that the smooth penalty function approximates $P(x) = \max(0, x)$ well for large γ. The right-hand side plot in Fig. 2 depicts the ‘zoomed’ graph of $P(x)=0.5x(1+\tanh(\gamma x))$ in a small area around the zero. One should always keep in mind that even for very large γ there will be a tiny area before the zero where the penalty function $P(x) = 0.5x(1 + \tanh(\gamma x))$ may be misleading with an incorrect maximum.



va

Fig. 2. Plots of $P(x) = 0.5x(1 + \tanh(\gamma x))$ for two γ values.

Solution of this approximation of the SCoTLASS problem (17)–(18) can be given as m consequent ascent gradient vector flows, each of them defined on an unit sphere in $\boldsymbol{R}^p$ and orthogonal to all preceding principal LASSO-components. Thus the loadings for the first $m(\leqslant p)$ principal LASSO-components $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ of the $p \times p$ correlation matrix $\mathbf{R}$ can be computed as solutions of m consequent initial value problems for the following vector ODEs:

$$\frac{\mathrm{d}\mathbf{a}_k}{\mathrm{d}t} = \pi_k \nabla_{F_\mu}(\mathbf{a}_k), \tag{19}$$

where the gradient of the function to be maximized (17) is given by the following $p \times 1$ vector:

$$\nabla_{F_\mu}(\mathbf{a}_k) = \mathbf{R}\mathbf{a}_k - 0.5\mu[1 + \tanh(\gamma y_k) + \cosh^{-2}(\gamma y_k)(\gamma y_k)]\mathbf{z}_k, \tag{20}$$

with

$$y_k = \mathbf{a}_k^{\mathrm{T}} \tanh(\gamma \mathbf{a}_k) - t \tag{21}$$

and

$$\mathbf{z}_k = \tanh(\gamma \mathbf{a}_k) + \mathrm{diag}(\cosh^{-2}(\gamma \mathbf{a}_k))(\gamma \mathbf{a}_k), \tag{22}$$

where for a $p \times 1$ vector $\mathbf{x}$ we denote by $\mathrm{diag}(\mathbf{x})$ the $p \times p$ diagonal matrix with the vector $\mathbf{x}$ on its main diagonal.

The m consequent ODEs (19) start with an appropriate initial value $\mathbf{a}_{k,\mathrm{in}}$ with $\|\mathbf{a}_{k,\mathrm{in}}\|_2 = 1$ for $k = 1, 2, \ldots, m$, see also the next section. The projector π_k is defined in (15).

First order necessary conditions for the critical points of the problem (17)–(18) can be obtained by simply making zero the right-hand side of the ODEs (19).

If $\mu > 0$ is interpreted as a Lagrange multiplier, then

$$y_k(1 + \tanh(\gamma y_k)) \leqslant 0, \tag{23}$$

is a necessary condition for a critical point of the SCoTLASS problem (17)–(18). The inequality (23) can be true only if $y_k < 0$ or $y_k = 0$, i.e.:

$$\mathbf{a}_k^{\mathrm{T}} \tanh(\gamma \mathbf{a}_k) \leqslant t. \tag{24}$$

2.4. Convergence matters

The ODEs (19) have continuously differentiable right-hand side and are defined on compact sets (manifolds). Thus for any initial value $\mathbf{a}_{k,\mathrm{in}}$ with $\|\mathbf{a}_{k,\mathrm{in}}\|_2 = 1$ for $k = 1, 2, \ldots, m$ Eq. (19) have unique solutions $\mathbf{a}_k(t)$ defined for all $t \in \boldsymbol{R}$. The smoothness of the objective function (17) and the compactness of the constrained manifolds (18) ensure the convergence of $\mathbf{a}_k(t)$ to a connected component of the set of critical points of (17) as $t \to \infty$ (e.g. Helmke and Moore, 1994). Not all of these critical points are solutions of the initial maximization problem (1)–(3), e.g. some of them are its minima. In addition there may be multiple local maxima/minima.

If $\mathbf{R}$ has distinct eigenvalues then the critical points of (17) are isolated and every solution $\mathbf{a}_k(t)$, $k = 1, 2, \ldots, m$ of (19) converges to one of them as $t \to \infty$ (Helmke and Moore, 1994). Note that if $\mathbf{a}_k$ defines a principal component, the vector $-\mathbf{a}_k$ also defines that principal component. Thus we have 2^m $p \times m$ matrices composed by these admissible $\mathbf{a}_k$, $k = 1, 2, \ldots, m$ for which the variance (1) has the same maximum.

3. Numerical results

In this section, we report some of our numerical experiments with Eq. (19). The computations are carried out by MATLAB 5.3 on a PC DELL 8100 under WINDOWS 2000. We use *ode15s* from the MATLAB ODE suite (MATLAB, 1999; Shampine and Reichelt, 1997) as the integrator for the initial value (Cauchy) problems. The code *ode15s* is a quasi-constant step size implementation of the Klopfenstein–Shampine family of the numerical differential formulas for stiff systems.

In our experiments, the tolerance for absolute error is set at 10^{-6} and for relative error—at 10^{-4}. This criterion is used to control the accuracy in following the solution path. We have experimented with many tests where the problem data are generated randomly. Because of the global convergence property of our method, all tests have similar dynamical behavior. The output values at time interval [0, 10] are examined. The integration terminates automatically when the relative improvement of the objective function between two consecutive output points is less than 10^{-4}, indicating a local maximizer has been found. We display all numbers only with four digits. All codes used in this experiment are available upon request.

The numerical illustrations hereafter are based on the correlation matrix for Jeffers' pitprop data (Jolliffe et al., 2003) (Table 1). In all reported experiments $\mu = 800$ in (17) and $\gamma = 1000$.

The initial (starting) value $\mathbf{a}_{1,\text{in}}$ for the ODE (19) with $k = 1$ is produced as follows. First, an $p \times 1$ vector is generated with random components each of them uniformly distributed in $[-0.5, 0.5]$, and, then, it is normalized to unit length. Solve the first ODE and let $\mathbf{a}_{1,\text{out}}$ denote its solution. Compose the projector $\mathbf{I}_p - \mathbf{a}_{1,\text{out}}\mathbf{a}_{1,\text{out}}^{\mathrm{T}}$ and apply it to an $p \times 1$ vector generated with random components each of them uniformly distributed in $[-0.5, 0.5]$; normalize the resulting vector to unit length and this is $\mathbf{a}_{2,\text{in}}$, etc.

Making use of such random initial values, the SCoTLASS problem has been solved in Jolliffe et al. (2003) for the Jeffers' pitprop data for several different values of the tuning parameter t. It has been noticed (see Jolliffe et al., 2003, Remark 4) that jumps occur in some LASSO-component loadings. It seems very likely to us that this is caused by the random starts used to solve the problem for different t. The feasible set of the SCoTLASS problem is a not connected subset of the unit sphere in $\boldsymbol{R}^p$. The random starts are simply vectors anywhere on the unit sphere. The exterior penalty function forces the starting vector to move towards one of the connected components of the feasible set. Obviously, it is not guaranteed that the same connected component will attract the solutions for different t starting with different (or even with same) random starts.

In order to avoid such unexpected jumps it seems reasonable to look for some appropriate rational starts. As it was mentioned before, the geometry of the feasible set is very helpful to achieve this. It suggests that the solution of the SCoTLASS problem (1)–(3) can be found by solving the problem in each component and picking the greatest maximum. For small values of p this is quite a reasonable approach. In this case random starts are not needed. Instead, the directional unit vectors $\mathbf{e}_k \in \boldsymbol{R}^p$ can be used as initial values $\mathbf{a}_{k,\text{in}}$ for the ODE (19).

Another possible way to avoid such unexpected jumps is to solve the problem in a sequential manner, i.e. the solution of the SCoTLASS problem for certain t_1 is used as a starting point of the algorithm to solve the problem for another $t_2 (< t_1)$. For comparison

Table 1
Loadings for SCoTLASS for four values of *t* based on the correlation matrix for Jeffers' pitprop data

Technique	Variable	Component					
		(1)	(2)	(3)	(4)	(5)	(6)
SCoTLASS ($t = 2.25$)	x_1	0.558	0.085	−0.093	−0.109	−0.057	−0.012
	x_2	0.581	0.031	−0.086	−0.148	−0.074	−0.042
	x_3	0.000	0.646	−0.133	0.214	0.066	0.099
	x_4	0.000	0.654	−0.000	0.209	0.078	−0.126
	x_5	−0.000	0.000	0.408	−0.000	−0.243	−0.748
	x_6	0.001	0.213	0.522	−0.018	0.107	−0.040
	x_7	0.266	−0.000	0.382	0.000	0.120	−0.027
	x_8	0.103	−0.097	0.000	0.584	−0.124	0.185
	x_9	0.372	−0.000	−0.000	0.022	−0.140	0.053
	x_{10}	0.364	−0.153	0.000	0.214	0.296	−0.000
	x_{11}	−0.000	0.100	−0.000	0.000	−0.878	0.168
	x_{12}	−0.000	0.240	−0.001	−0.698	0.048	0.180
	x_{13}	−0.000	0.023	−0.618	−0.024	0.009	−0.559
SCoTLASS ($t = 2.00$)	x_1	0.623	0.036	−0.057	−0.047	−0.035	0.041
	x_2	0.647	0.000	−0.065	−0.082	−0.045	0.001
	x_3	0.000	0.657	−0.169	0.121	0.070	0.091
	x_4	0.000	0.676	−0.000	0.110	0.040	−0.096
	x_5	−0.000	0.001	0.012	0.000	−0.294	−0.916
	x_6	0.000	0.239	0.474	0.000	0.011	−0.141
	x_7	0.137	0.000	0.572	0.020	0.039	−0.031
	x_8	0.001	−0.022	0.000	0.753	−0.085	0.084
	x_9	0.332	−0.000	−0.000	0.001	−0.084	0.001
	x_{10}	0.253	−0.091	0.001	0.310	0.289	−0.087
	x_{11}	0.000	0.073	−0.000	−0.000	−0.893	0.264
	x_{12}	−0.000	0.196	−0.000	−0.549	0.081	0.062
	x_{13}	0.000	0.000	−0.642	−0.000	0.027	−0.177
SCoTLASS ($t = 1.75$)	x_1	0.663	−0.000	−0.000	−0.024	−0.031	0.004
	x_2	0.683	−0.001	−0.000	−0.040	−0.019	0.001
	x_3	0.000	0.642	−0.193	0.000	−0.001	0.147
	x_4	0.000	0.701	−0.001	0.000	−0.006	−0.000
	x_5	−0.000	0.000	0.000	0.000	−0.218	−0.905
	x_6	0.000	0.293	0.187	0.000	0.000	−0.321
	x_7	0.001	0.106	0.651	−0.000	0.043	−0.001
	x_8	0.001	−0.000	0.000	0.734	−0.136	0.014
	x_9	0.284	−0.000	0.000	0.000	−0.001	0.000
	x_{10}	0.111	−0.000	0.001	0.387	0.301	−0.026
	x_{11}	0.000	0.000	−0.000	−0.000	−0.915	0.199
	x_{12}	−0.000	0.001	−0.000	−0.556	0.033	0.000
	x_{13}	0.000	−0.000	−0.710	0.001	0.040	−0.126
SCoTLASS ($t = 1.50$)	x_1	0.701	−0.000	−0.000	−0.001	−0.000	−0.000
	x_2	0.708	−0.000	−0.000	−0.001	−0.000	0.000
	x_3	0.000	0.698	−0.069	0.003	0.000	0.001
	x_4	0.000	0.712	−0.000	0.001	−0.000	0.001
	x_5	−0.000	0.000	0.000	0.000	−0.868	−0.165

Table 1 (*continued*)

Technique	Variable	Component					
		(1)	(2)	(3)	(4)	(5)	(6)
	x_6	0.000	0.082	0.587	−0.031	0.000	−0.009
	x_7	0.001	0.000	0.806	−0.001	0.001	−0.000
	x_8	0.000	−0.000	0.000	0.768	0.000	−0.166
	x_9	0.083	0.000	0.000	0.001	−0.000	−0.000
	x_{10}	0.001	−0.000	0.030	0.638	−0.000	0.199
	x_{11}	0.000	0.000	−0.000	−0.000	0.150	−0.952
	x_{12}	−0.000	0.001	−0.000	−0.051	0.000	−0.000
	x_{13}	0.000	0.000	−0.000	−0.000	0.474	0.000

we solve the SCoTLASS problem for the Jeffers' pitprop data for the same values of the tuning parameter t as in Jolliffe et al. (2003). These results can be compared to the solutions found in Table 4 (Jolliffe et al., 2003).

We start the SCoTLASS algorithm with the PCA solution for the Jeffers' pitprop data given in Table 2 (Jolliffe et al., 2003). When we want to solve the SCoTLASS problem for $t = 2.25$, instead of starting the algorithm with a random initial value, we start from the already found PCA solution. Similarly, this solution obtained for $t = 2.25$ is used as a starting point to solve the SCoTLASS problem with $t = 2.00$ and so on. The sequential solutions change more gradually in the sense that the orientation of the principal components is inherited from step t_1 to the next step $t_2 (< t_1)$, compare Tables 1 and 4 (Jolliffe et al., 2003). Each sequential run requires considerably less CPU time (2–3 times) than the corresponding run with arbitrary random start. Another advantage of the sequential strategy is that it results in global maxima (or at least quite near) and there is no need to re-start the algorithm many times. Still there is a danger with this approach. Suppose, for certain t, the components $\mathbf{c}_3$ and $\mathbf{c}_4$ explain nearly equal variances. If the numerical solution for $\mathbf{c}_3$ is a local minimum, then the solution for $\mathbf{c}_4$ may produce larger variance. Then the corresponding columns of loadings exchange their places, which may spoil the sequential process and is followed by misleading results. Random starts are required to avoid the local solution.

In Table 2 are listed the values of the normalized VARIMAX function as a measure of simplicity, the variances and the number of zeros for each of the obtained SCoTLASS solutions. These results are very similar to our findings in Table 4 (Jolliffe et al., 2003). The new solution for $t = 1.50$ has considerably more "clearer" zeros.

4. Concluding remarks

In this work we consider and solve the PCA based on the LASSO (1)–(3) called SCoTLASS. The LASSO inequality constraint is tackled by introducing an exterior penalty function. The transformed objective function is then maximized subject to equality constraints making use of the projected gradient approach which follows precisely the geometry of the constraints (Chu and Trendafilov, 2001). This was demonstrated with several numerical examples where both the LASSO inequality and orthonormality equality constraints are perfectly satisfied.

Table 2
Simplicity factor, variance, cumulative variance and number of zero loadings for individual components in PCA, RPCA, and SCoTLASS for four values of t, based on the correlation matrix for Jeffers' pitprop data

Technique	Measure	Component					
		(1)	(2)	(3)	(4)	(5)	(6)
SCoTLASS	Simplicity factor (varimax)	0.190	0.311	0.208	0.308	0.574	0.366
($t = 2.25$)	Variance (%)	26.7	17.2	15.9	9.7	8.9	6.7
	Cumulative variance (%)	26.7	43.9	59.7	69.4	78.3	85.0
	Number of zero loadings	6	3	5	3	0	1
SCoTLASS	Simplicity factor (varimax)	0.288	0.350	0.272	0.373	0.613	0.689
($t = 2.00$)	Variance (%)	23.1	16.5	14.7	11.6	8.8	8.3
	Cumulative variance (%)	23.1	39.6	54.3	65.9	74.7	83.1
	Number of zero loadings	7	5	5	4	0	0
SCoTLASS	Simplicity factor (varimax)	0.370	0.370	0.389	0.359	0.689	0.657
($t = 1.75$)	Variance (%)	19.6	16.0	13.2	13.1	9.1	8.8
	Cumulative variance (%)	19.6	35.6	48.7	61.8	70.8	79.7
	Number of zero loadings	7	7	7	7	1	3
SCoTLASS	Simplicity factor (varimax)	0.451	0.451	0.503	0.473	0.586	0.809
($t = 1.50$)	Variance (%)	16.1	14.9	13.9	12.2	8.8	8.7
	Cumulative variance (%)	16.1	31.0	44.9	57.0	65.9	74.6
	Number of zero loadings	8	9	9	3	9	6

An open question related to the SCoTLASS problem is how to choose the tuning parameter t such that the solution exhibits reasonable interpretability, while still containing considerable part of the sample variance (1).

Acknowledgements

We are grateful to the Associate Editor for helpful comments on the first draft of this paper.

References

Aoki, M., 1971. Introduction to Optimization Techniques: Fundamentals and Applications of Nonlinear Programming. The Macmillan Company, New York.

Brockett, R.W., 1989. Least squares matching problems. Linear Algebra Appl. 122/123/124, 761–777.

Chen, T., Amari, S.I., Lin, Q., 1998. A unified algorithm for principal and minor components extraction. Neural Networks 11, 385–390.

Chu, M.T., Trendafilov, N., 1998. On a differential equation approach to the weighted orthogonal Procrustes problem. Statist. Comput. 8, 125–133.

Chu, M.T., Trendafilov, N.T., 2001. The orthogonally constrained regression revisited. J. Comput. Graphical Statist. 10, 746–771.

Edelman, A., Arias, T., Smith, S.T., 1998. The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. 20 (2), 303–353.

Golub, G.H., Van Loan, Ch.F., 1991. Matrix computations, 2nd ed.. The John Hopkins University Press, Baltimore, London.

Helmke, U., Moore, J.B., 1994. Optimization and Dynamical Systems. Springer, London.

Hirsch, M.W., Smale, S., 1974. Differential Equations, Dynamical Systems, and Linear Algebra. Academic Press, London.

Jolliffe, I.T., 2002. Principal Component Analysis. 2nd ed. Springer, New York.

Jolliffe, I.T., Uddin, M., 2000. The simplified component technique–an alternative to rotated principal components. J. Comput. Graphical Statist. 9, 689–710.

Jolliffe, I.T., Trendafilov, N.T., Uddin, M., 2003. A modified principal component technique based on the LASSO. J. Comput. Graphical Statist. 12, 531–547.

Mahony, R.E., Helmke, U., Moore, J.B., 1996. Gradient algorithms for principal component analysis. J. Australian Math. Soc. B 37, 430–450.

MATLAB (1999). Using MATLAB, Version 5, The MathWorks Inc., Natick, MA.

Osborne, M.R., Presnell, B., Turlach, B.A., 2000a. A new approach to variable selection in least squares problems. IMA J. Numer. Anal. 20, 389–403.

Osborne, M.R., Presnell, B., Turlach, B.A., 2000b. On the LASSO and its dual. J. Comput. Graphical Statist. 9, 319–337.

Shampine, L.F., Reichelt, M.W., 1997. The MATLAB ODE suite. SIAM J. Sci. Comput. 18, 1–22.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. J. Roy. Statist. Soc. B 58, 267–288.

Trendafilov, N.T., 2001. Principal components simplifying the ORTHOMAX criterion. International Meeting of the Psychometric Society (IMPS-2001), 15–19 July, Osaka, Japan.

Turlach, B.A., Venables, W.N., Wright, S.J., 2001. Simultaneous variable selection. Technical Report, http://www.maths.uwa.edu.au/∼berwin/psfiles/tvw.pdf